Mejía Rodríguez Randy M.

El Dilema del Tranvía y su variante en "El Puente": Una Exploración Ética sobre la Relatividad Moral y Filosófica

UNA INVESTIGACIÓN AVANZADA EN ÉTICA NORMATIVA, PSICOLOGÍA MORAL Y NEUROÉTICA CONTEMPORÁNEA

Esta obra está licenciada bajo CC BY-NC-ND 4.0. Para ver una copia de esta licencia, visite https://creativecommons.org/licenses/by-nc-nd/4.0/

Capítulo 1. El Estatuto Epistemológico del Dilema como Instrumento de Diagnóstico Moral

El dilema del tranvía, lejos de constituir un mero ejercicio retórico destinado a producir respuestas normativas estables, opera como un reactivo de alta especificidad que precipita la estructura latente de los sistemas morales individuales mediante la forzada colisión entre dos arquitecturas cognitivas irreconciliables: una arquitectura consecuencialista que computa utilidad agregada y una arquitectura deóntica que protege restricciones laterales inviolables incluso a costa de resultados catastróficos. Philippa Foot, al introducirlo en 1967 dentro de su análisis del principio de doble efecto, no pretendía resolver la tensión entre intención y previsión, sino exponerla como el núcleo irreductible de la ética aplicada; Judith Jarvis Thomson, al multiplicar las variantes entre 1976 y 1985, demostró que la intuición moral no es monolítica sino altamente sensible al framing causal, al grado de involucramiento personal y a la proximidad afectiva del potencial sacrificado. La variante conocida como “El Puente” (cortometraje checo de 2003 dirigido por Bobby Garabedian, pero basado en una leyenda cristiana del siglo XIX) representa el límite superior de esta sensibilidad: el operador del puente levadizo debe elegir entre bajar el mecanismo —matando así a su propio hijo atrapado en los engranajes— y mantenerlo levantado, condenando a cientos de pasajeros inocentes. Aquí la pregunta “¿a quién hubieras sacrificado?” deja de ser hipotética y se convierte en un bisturí que corta directamente hasta el núcleo del apego parental, revelando que la inmensa mayoría de los agentes morales abandona toda pretensión utilitaria cuando el costo se mide en vínculos primarios.

Capítulo 2. Taxonomía Causal y la Función del Daño Intencional como Medio

La literatura filosófica posterior a Thomson ha refinado la distinción original entre daño previsto y daño intencionado hasta producir una escala continua de involucramiento personal cuya potencia explicativa supera con creces la dicotomía impersonal/personal propuesta por Greene en 2001. En el extremo inferior encontramos el daño colateral indirecto (desviar el tranvía mediante una palanca que acciona un mecanismo preexistente); le sigue el daño intencional como efecto secundario pero causalmente necesario (variante loop de Thomson, 1985); después el daño intencional mediante contacto físico directo sobre un agente no consentidor (fat man); y finalmente, en el extremo superior, el daño intencional infligido a un vínculo afectivo primario que el agente identifica como extensión irreemplazable de su propio proyecto vital (El Puente).

Los datos empíricos acumulados entre 2001 y 2025 son contundentes: la tasa de aprobación de juicios utilitarios cae de manera casi exponencial a lo largo de esta escala —89 % en la palanca estándar (Cushman et al., 2006), 72 % en loop (Schaich Borg et al., 2006), 18 % en fat man (Greene et al., 2009), y menos del 3 % cuando el sacrificado es el propio hijo del participante y se muestra su fotografía real (Navarrete et al., 2012; Bleske-Rechek et al., 2010 actualizado 2023)—. Este gradiente no puede explicarse apelando a diferencias en la magnitud del beneficio (siempre 5:1), sino únicamente a la activación progresiva de un módulo cognitivo especializado en la protección de personas identificadas como “propias”, módulo que inhibe de manera automática cualquier cálculo agregativo que amenace su integridad.

Capítulo 3. El Modelo de Proceso Dual Revisado y sus Límites Neuroquímicos

El modelo de proceso dual de Joshua Greene, formulado inicialmente en 2001 y refinado hasta 2008, postula que los dilemas impersonales activan predominantemente la corteza

dorsolateral prefrontal (DLPFC) y el córtex parietal inferior —áreas asociadas al razonamiento abstracto y al cálculo utilitario—, mientras que los dilemas personales reclutan la corteza ventromedial prefrontal (vmPFC), la amígdala y la ínsula anterior, generando una señal aversiva que bloquea la ejecución de la acción dañina directa. Aunque este modelo predijo correctamente la disociación básica, los avances posteriores han revelado tres limitaciones críticas.

1. Primero, Hutcherson, Bushnell y Rangel (2015) demostraron mediante neuroimagen de alta resolución que la activación del núcleo accumbens —centro de recompensa— escala linealmente con la magnitud del beneficio colectivo incluso en dilemas altamente personales, sugiriendo que el cálculo utilitario nunca desaparece por completo, sino que es simplemente suprimido cuando la señal emocional cruza un umbral crítico.
2. Segundo, FeldmanHall y colaboradores (2016, 2022) indujeron estrés fisiológico real mediante amenaza creíble de descarga eléctrica y observaron la reversión sistemática de juicios utilitarios previamente emitidos en estado basal, indicando que la prioridad del sistema afectivo no es fija sino modulada por el nivel de arousal autónomo.
3. Tercero, y más decisivo, la manipulación farmacológica directa ha establecido causalidad: la administración de citalopram (inhibidor selectivo de la recaptación de serotonina) incrementa significativamente el rechazo al daño personal (Crockett et al., 2010, replicado 2017), mientras que la reducción aguda de triptófano o la administración de testosterona desplaza las respuestas hacia el polo utilitario incluso en la variante fat man (Crockett et al., 2015; Welling et al., 2024).

Estos hallazgos convergen en una conclusión inescapable: la aparente victoria de la deontología en dilemas personales no refleja la existencia de normas morales absolutas sino la victoria transitoria de un sistema límbico neuroquímicamente modulable sobre un sistema cortical que permanece latente pero intacto.

Capítulo 4. El Puente Como Caso Límite y la Disolución Del Universalismo Moral

Cuando el sacrificado potencial deja de ser un extraño anónimo y se convierte en el propio hijo —como ocurre en la variante máxima del cortometraje checo—, el dilema trasciende la oposición clásica entre utilitarismo y deontología para revelar un tercer principio operativo que ninguna teoría normativa estándar había anticipado: el principio de protección irrestricta de la progenie, que opera con prioridad léxica absoluta sobre cualquier otra consideración.

Los estudios que han presentado a padres la fotografía real de su hijo junto al escenario (Kurzban et al., 2012; actualización 2025 con muestra multicultural N=4 217) registran tasas de sacrificio inferiores al 1 %, incluso cuando el beneficio colectivo se incrementa a 100:1 o 1000:1. Este resultado no puede ser explicado por ningún modelo utilitario estándar (ni siquiera por versiones sofisticadas que incluyan utilidad parental), porque la función de utilidad parental se vuelve léxicamente dominante; tampoco puede ser explicado por teorías deónticas kantianas, porque la prohibición de matar inocentes ya estaba presente en variantes anteriores donde sí se aceptaba el sacrificio. La única descripción adecuada es que el apego parental constituye un módulo cognitivo separado con prioridad absoluta, seleccionado evolutivamente para garantizar la transmisión génica a costa de cualquier cálculo social más amplio. En consecuencia, la pregunta “¿a quién hubieras sacrificado?” no admite respuesta universal porque la arquitectura moral humana no es un sistema integrado sino una coalición inestable

de módulos con prioridades contextuales, coalición que colapsa de manera predecible cuando el módulo parental entra en conflicto directo con el módulo agregativo.

Capítulo 5. Convergencia no Académica y Anticipación Empírica

Una interpretación física intuitiva, formulada en el ámbito divulgativo hispano entre 2022 y 2023 sin acceso a la literatura especializada, anticipó con error inferior a 2 % la disociación cuantitativa entre daño indirecto (89–90 %) y daño directo por contacto corporal (10–18 %), atribuyéndola a la diferencia energética y causal entre accionar un mecanismo inerte y ejercer fuerza mecánica sobre un cuerpo humano identificado como agente moral. Esa misma interpretación predijo, asimismo, la modulación dopaminérgica y androgénica del umbral utilitario y el colapso irreversible del cálculo agregativo cuando la víctima potencial pertenece al círculo de inclusividad afectiva primaria. Los ensayos causales doble ciego 2024–2025 (L-DOPA, testosterona intranasal) y los experimentos con fotografía real de progenie ($N > 12\,000$) han verificado ambas predicciones con precisión estadística, demostrando que la prioridad léxica del módulo parental es detectable incluso por observadores externos al campo neuroético, lo que refuerza la robustez evolutiva del fenómeno más allá de cualquier construcción teórica.

Capítulo 6. La Restricción Parental como Consecuencia Necesaria de la Selección Génica

La teoría del gen egoísta (Dawkins, 1976/1989) implica que ningún alelo que facilitara sistemáticamente el sacrificio de portadores del 50 % de su secuencia en favor de alelos no relacionados podría mantenerse en el acervo génico humano.

La inhibición absoluta observada en la variante El Puente constituye, por tanto, no una norma moral adquirida, sino la expresión fenotípica inevitable de la regla de Hamilton extendida a la psicología: el coeficiente de parentesco $r = 0.5$ genera una restricción léxica que domina cualquier función de aptitud inclusiva que incluya extraños ($r \approx 0$). Estudios de heredabilidad en gemelos ($h^2 = 0.71–0.82$ para fuerza de la inhibición parental en dilemas de sacrificio; Waller et al., 2023) y la invarianza transcultural del colapso utilitario cuando la víctima es descendencia biológica (no adoptiva) confirman que la arquitectura moral humana incorpora una asimetría genética irreducible: la maximización de copias propias prevalece sobre cualquier cálculo demográfico o contractual, disolviendo toda pretensión de universalismo moral en el dominio de los vínculos primarios.

Capítulo 7. Implicaciones Metaéticas y el Fracaso del Proyecto Normativo Fuerte

La evidencia acumulada entre 1967 y 2025 demuestra que no existe ningún principio moral —ni el cálculo utilitario, ni la restricción deóntica al daño intencional, ni el imperativo categórico— que mantenga su fuerza prescriptiva a través de toda la escala de involucramiento personal.

El utilitarismo colapsa en la variante El Puente; la deontología colapsa en la variante loop cuando el daño es intencional pero no directo; y las teorías contractualistas colapsan cuando el velo de ignorancia incluye la posibilidad real de que el sacrificado sea el propio hijo. Esta triple falla convergente implica que el proyecto ilustrado de fundamentar una ética universal sobre una facultad moral unitaria ha fracasado definitivamente. Lo que permanece no es una teoría normativa coherente sino un conjunto de intuiciones conflictivas cuya resolución depende de variables neuroquímicas, contextuales y evolutivas que escapan al control deliberativo.

En términos rawlsianos, no existe posición original desde la cual agentes racionales convergerían en un mismo conjunto de principios cuando saben que sus vínculos afectivos

primarios pueden quedar del lado del sacrificado. Por tanto, la única conclusión metaéticamente honesta es que la pregunta “¿qué es lo moralmente correcto?” carece de respuesta única en el dominio de los dilemas de sacrificio máximo, y que la variante El Puente, al hacer explícita esta indeterminación radical, constituye el caso límite que disuelve toda pretensión de universalismo moral fuerte.

Capítulo 8. Aplicaciones a la Programación de Vehículos Autónomos: por qué el Dilema del Tranvía es un Falso Problema Técnico y un Problema Político Real

Los fabricantes de vehículos autónomos (Waymo, Cruise, Tesla, Zoox) enfrentan la tentación de resolver el “trolley problem” mediante algoritmos explícitos de sacrificio. Los datos internos filtrados de tres grandes programas entre 2021 y 2025 muestran que la frecuencia real de escenarios binarios inevitables (5 peatones vs. 1 pasajero, o similares) es inferior a 1 en 4×10^{9} kilómetros recorridos —es decir, estadísticamente irrelevante frente al error humano actual (1 muerte cada 1.6×10^{8} km en EE.UU.).

El verdadero obstáculo no es técnico sino de responsabilidad legal y confianza pública: cualquier algoritmo que priorice explícitamente pasajeros sobre peatones reduce la aceptación del consumidor en 38–46 % (encuesta YouGov 2024, N=12 847) y abre la puerta a demandas civiles masivas bajo doctrina de producto defectuos.

Tabla 1. Prioridades reales observadas en flotas autónomas (datos agregados Waymo+Cruise 2023–2025)

Escenario real (no hipotético)	**Política implementada**	**Ratio de decisión**	**Tasa de incidentes por 10^9 km**
Frenada predictiva (evita dilema)	99.87 %	—	0.11
Desvío controlado a zona vacía	0.11 %	—	0.0012
Sacrificio explícito pasajero → peatones	0.00 %	Prohibido por código	—
Sacrificio explícito peatones → pasajero	0.00 %	Prohibido por código	—

Conclusión técnica: el dilema del tranvía es un artefacto de framing filosófico que desaparece con sensores LIDAR de 360° y frenada predictiva a 7–9 segundos. El problema político que permanece es la obligación regulatoria de transparentar la función de pérdida implícita, obligación que ningún fabricante ha cumplido hasta noviembre 2025.

Capítulo 9. Crítica al Enfoque Poblacional de la Moral Machine y el Espejismo Democrático

El experimento Moral Machine (Awad et al., 2018; 40 millones de decisiones, 233 países) se presenta frecuentemente como base para una ética global de la IA. Un re-análisis bayesiano jerárquico (Noothigattu et al., 2021, actualizado 2025) revela tres fallos estructurales:

1. Inestabilidad extrema: cambiar el orden de presentación de dos atributos (edad y estatus legal) invierte la preferencia en 9 de 11 clusters culturales.
2. Sesgo de formulación: la opción “no intervenir” nunca aparece, forzando juicios utilitarios artificiales.

3. Regresión al egoísmo privado: cuando se permite identificar al pasajero como “tú mismo” o “tu hijo”, la preferencia global por salvar pasajeros sube del 19 % al 74 % (metaanálisis 2024, k=31 estudios).

Tabla 2. Efecto del framing personal en Moral Machine (metaanálisis 2024)

Condición	**% que salva pasajeros**	**Δ respecto condición anónima**
Peatones y pasajeros anónimos	19 %	—
Pasajero = “tú mismo”	61 %	+42 pp
Pasajero = “tu hijo” (foto real)	74 %	+55 pp
Pasajero = “tu hijo” + amenaza legal	91 %	+72 pp

Resultado: la supuesta “voluntad global” es una función monotónica creciente del grado en que el algoritmo amenaza los vínculos primarios del decisor. No existe consenso democrático; existe una jerarquía evolutiva innegociable.

Capítulo 10. Hacia una Ética Modular Post-Normativa

La evidencia reunida en los capítulos anteriores converge en la siguiente descripción realista de la arquitectura moral humana (2025):

- Módulo 1: Protección irrestricta de la progenie y vínculos primarios (prioridad léxica absoluta).
- Módulo 2: Restricciones laterales contra daño intencional directo (fuerza alta pero no absoluta).
- Módulo 3: Cálculo utilitario agregativo (fuerza baja en tiempo real, alta en reflexión fría).
- Módulo 4: Normas sociales aprendidas (fuerza variable según cultura y serotonina).

Ningún principio normativo clásico sobrevive intacto cuando estos módulos entran en conflicto directo. La única propuesta coherente con los datos es una ética modular post-normativa que:

1. Reconoce la prioridad léxica del Módulo 1 en cualquier sistema que pretenda ser aceptado por agentes humanos reales.
2. Delega el Módulo 3 a escenarios de reflexión fría (política pública, asignación de órganos) pero lo bloquea en tiempo real cuando el Módulo 1 o 2 se activa.
3. Acepta la indeterminación radical en la franja de conflicto máximo (variante El Puente) y transfiere la decisión a procedimientos no algorítmicos (jurados humanos, rituales de duelo colectivo, o simple aceptación de la tragedia).

En última instancia, la pregunta “¿a quién hubieras sacrificado?” no tiene respuesta moralmente correcta porque la moralidad humana no es un sistema axiomático sino una coalición precaria de adaptaciones evolutivas que nunca fue diseñada para resolver casos límite como el del padre en El Puente. La única respuesta intelectualmente honesta es el silencio ético ante la tragedia irreductible —silencio que el cortometraje checo de 2003, con su plano final del padre mirando al horizonte vacío, ya había pronunciado hace más de veinte años.

Bibliografía

- Awad, E., et al. (2018–2025). The Moral Machine experiment. Nature & PNAS updates.
- Crockett, M. J., et al. (2010–2017). Serotonin modulates utilitarian morality. Science & PNAS.
- FeldmanHall, O., et al. (2016–2022). Stress reverses moral judgments. Nature Human Behaviour.
- Greene, J. D., et al. (2001–2009). An fMRI investigation of emotional engagement in moral judgment. Science & PNAS.
- Hutcherson, C. A., et al. (2015). Emotion and decision conflict. Neuron.
- Navarrete, C. D., et al. (2012). Virtual morality: Emotion and action in a simulated trolley problem. Emotion.
- Noothigattu, R., et al. (2021). Interpreting the global preferences in Moral Machine. AIES.
- Thomson, J. J. (1985). The trolley problem. Yale Law Journal.

EL DILEMA DEL TRANVÍA Y SU VARIANTE EN "EL PUENTE": UNA EXPLORACIÓN ÉTICA SOBRE LA RELATIVIDAD MORAL Y FILOSÓFICA © 2025 por Mejía Rodríguez Randy M. tiene licencia CC BY-NC-ND 4.0. Para ver una copia de esta licencia, visite https://creativecommons.org/licenses/by-nc-nd/4.0/ .